Jointly Learning Heterogeneous Features for RGB-D Activity Recognition

Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang

Abstract—In this paper, we focus on heterogeneous features learning for RGB-D activity recognition. We find that features from different channels (RGB, depth) could share some similar hidden structures, and then propose a joint learning model to simultaneously explore the shared and feature-specific components as an instance of heterogeneous multi-task learning. The proposed model formed in a unified framework is capable of: 1) jointly mining a set of subspaces with the same dimensionality to exploit latent shared features across different feature channels, 2) meanwhile, quantifying the shared and feature-specific components of features in the subspaces, and 3) transferring feature-specific intermediate transforms (i-transforms) for learning fusion of heterogeneous features across datasets. To efficiently train the joint model, a three-step iterative optimization algorithm is proposed, followed by a simple inference model. Extensive experimental results on four activity datasets have demonstrated the efficacy of the proposed method. A new RGB-D activity dataset focusing on human-object interaction is further contributed, which presents more challenges for RGB-D activity benchmarking.

Index Terms—heterogeneous features learning, RGB-D activity recognition, action recognition

1 Introduction

The emergence of low-cost depth sensors (e.g., the Microsoft Kinect) opens a new dimension to address the challenges of human activity recognition. Compared to the conventional use of RGB videos, the information from depth channel is insensitive to illumination variations, invariant to color and texture changes, and more importantly reliable for body silhouette and skeleton (human posture) extraction [31]. Bearing on these merits, it is believed that the introduced depth information can greatly light up the research of human activity analysis [12], [24], [36].

Nevertheless, using depth alone has limitations in distinguishing human activities and object context in challenging cases [39], [55]. Depth information (e.g. captured by existing Kinect device) often suffers from low spatial resolution and low depth precision. Moreover, the depth information is usually weak in capturing the important appearance information, such as object color and texture. These greatly limit the application of depth cameras on recognizing complex human activities with object and interactions, such as human-object interactions [10], [50] and fine grained activities [16], where the color appearance is also important.

In fact, there indeed exists a connection between the information from RGB and depth channel, which could be unveiled after certain transformation. In Figure 1, we show some visualization results of the HOG features extracted from RGB image patches and the corresponding depth patches. Albeit extracted from different channels, these HOG features still look similar for each of the activities. This suggests that depth channel is related to the RGB channels and the heterogeneous features extracted from different channels could share some (hidden) structures (e.g., the *gist* of looking down in *reading*, and cup-to-mouth in *drinking* as shown in Figure 1). However, despite the similarities in the visualized HOG features, there still exist differences between different channels; for instance, the RGB channel mainly captures the appearance (color) information, while the depth channel describes the geometry (shape) cues in depth. This suggests they indeed have their own characteristics in describing objects. Hence, learning RGB and depth features together should not only extract shared features that are robust and collaborative across feature channels but also exploit features complementary between different channels. However, the majority of existing RGB-D action recognition methods [5], [20], [30] neither seek to jointly learn the features extracted from RGB and depth channels simultaneously nor model their underlying connections.

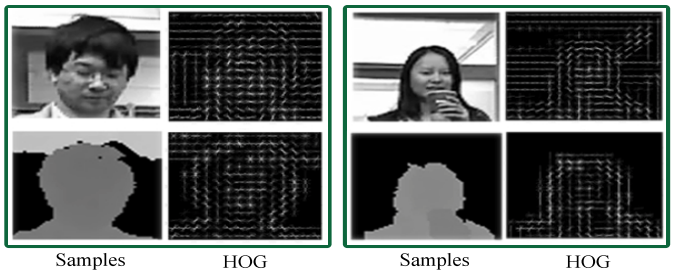


Fig. 1. Visualization of HOG features for two activity snapshots from RGB (gray) channel and depth channel, respectively. As shown, the HOG features from both channels of the same activity unveil similar "gist" structure of that activity, e.g., the"gist" of looking down in reading, and cup-to-mouth in drinking.

In order to effectively capture the connections among different

- *J.-F. Hu, W.-S. Zheng and J. Lai are with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. J.-F. Hu is also with the Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China. W.-S. Zheng is also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China. J. Lai is also with Guangdong Province Key Laboratory of Information Security, P. R. China. E-mail:hujianf@mail2.sysu.edu.cn, wszheng@ieee.org, and stsljh@mail.sysu.edu.cn*
- *J. Zhang is with the School of Science and Engineering (Computing), University of Dundee United Kingdom. E-mail:j.n.zhang@dundee.ac.uk*

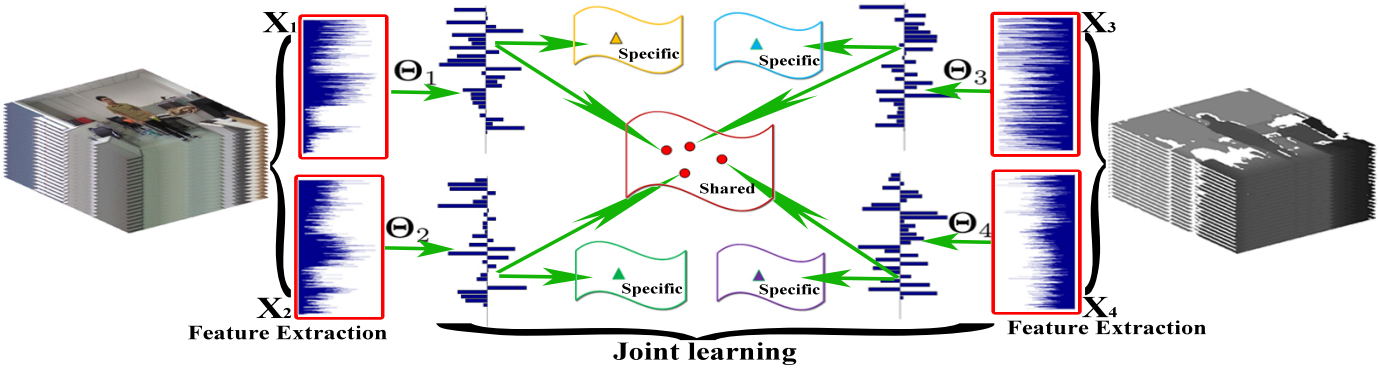


Fig. 2. A graphic illustration of our joint learning framework. In this framework, all the i-transforms (e.g. four i-transforms, $\{\Theta_i\}_{i=1,2,3,4}$) shared structures and specific structures are jointly learned for the purpose of recognition on RGB and Depth channels.

heterogeneous features, we propose a joint heterogeneous feature learning model for RGB-D activity recognition. In the proposed model, we learn a set of subspaces (one subspace for each heterogeneous feature type) such that features with different dimensionality can be compared, and their shared and specific components can be easily encoded. To achieve this, we introduce a linear projection matrix called the *intermediate transform* (i-transform) for each feature type, with the ability to control the dimensionality of each subspace. We then formulate our subspaces mining, shared and feature-specific components learning in the framework of multi-task learning. Therefore, the optimal solution for the i-transforms and shared-specific structures can be jointly derived, with the principles illustrated in Figure 2. Modeling in such a way can significantly improve the intrinsic structures learning among the features of different types and transfer knowledge between them. A three-step iterative optimization algorithm is proposed to find the optimal solution with a guaranteed convergence. We call the proposed model the **j**oint hete**r**ogeneous feat**u**res **le**arning (JOULE) model. Technically speaking, although efforts of exploring both shared and specific structures for classification are attempted in some of the existing multi-task learning methods [1], [2], [6], [54], our proposed model differs in that these methods assume that the features are homogeneous (the same type, e.g. word frequencies for text categorization) with the same dimensionality, thus not applicable for mining shared and feature-specific structures among heterogeneous features.

RGB-D training data in a target set are not always sufficient, in which case an auxiliary set is usually beneficial. To enable our model to handle this case, we further propose a transfer version of our JOULE, which is capable of effectively utilizing an auxiliary set. We assume that during learning, features of the same type from the auxiliary set and target set shares the same i-transform and can be jointly learned. Therefore, the knowledge transfer from auxiliary set to the target set could be achieved by the shared linear i-transforms, and subsequently enhance the recognition performance on the target set.

In addition to the aforementioned joint heterogeneous learning model, we present a variant of temporal pyramid Fourier features (TPF) developed in [39] so as to apply both the original feature signal and its gradient to implicitly encode human motions, which experimentally yield better performance than TPF on original feature signal only. And, in order to test the generalization performance of our method on 3D human-object interactions more extensively, we also contribute a new RGB-D activity dataset called *SYSU 3D HOI* activity set, which consists of 12 activity classes from 40 participants. Both this dataset and our codes will be available in http://isee.sysu.edu.cn/~hujianfang/ProjectJOULE.html.

In summary, the main contributions of our work are threefold: 1) a novel joint heterogeneous feature learning framework for RGB-D activity recognition, which is capable of learning hidden connections among heterogeneous features extracted from sequences of different channels; 2) a transfer RGB-D feature learning framework leveraging auxiliary datasets; 3) a new dataset collected for RGB-D human-object interaction recognition.

2 Related Work

Recently, recognizing human activities from low cost depth cameras has become a more and more important research direction with many applications including digital surveillance, virtual reality, human-computer interaction and Xbox One games etc. There are two emerging branches in activity recognition research: 1) depth-based representation, and 2) RGB-D based development. In this section, in addition to reviewing existing works of recognizing human activities captured by depth cameras, we further briefly describe the literature of learning heterogeneous features for generic visual recognition purpose, which is also relevant to ours.

Depth-based representation. On building depth-based representation, a straightforward way is to generalize the descriptors specially designed for RGB channel to depth channel for describing the shape geometry [18], [59]. For instance, Oreifej and Liu [28] extended the histogram of gradient (HOG) descriptor by constructing a histogram to capture the distribution of surface normal orientation in 4D space. Yang et al. [48] suggested that concatenating the normal vectors within a spatiotemporal depth sub-volume together can capture more informative geometric cues. [38] sought to explicitly encode the geometric cues by computing the number of points that follow in each sampled sub-volume. Lu et al. [21] directly investigated the relationship between sampled pixels in both actor and background regions. Most of these methods attempted to mine some discriminative local patterns for representing human activities without considering the holistic human poses, which has been demonstrated to be critical for describing complex human activities involving human-object interactions [10], [41], [50]. Due to the development of realtime human skeleton (3D posture) tracker from single depth image [31], human motions can be effectively captured by the positional dynamics of each individual skeletal joint [7], [13], [23],

[44] or the relationship of joint pairs [22], [27], [47] or even their combination [19], [52], [58]. Vemulapalli et al. [34] exploited 3D relative geometries among different body parts in the Lie algebra. In addition to the skeleton motions, local depth patterns are also found to be useful for discriminating complex activities with human-object interactions [39], [41]. Specifically, Wei et al. [41] presented a model to explicitly study the interactions of human and object. Koppula et al. [15] simultaneously modeled the human activities and object affordances in RGB-D videos with a structural support vector machine.

RGB-D based development. Depth does not necessarily mean discriminant. Albeit invariant to lighting changes, it does lose some useful information such as texture context, which is critical to distinguish some activities involving human-object interactions. Recent works also showed that the fusion of the RGB and depth sequences can largely improve the recognition of activities with object interactions [5], [15], [16], [20], [30], [41], [51], [55]. For instance, Zhao et al. [55] combined interest point based descriptors extracted from RGB and depth sequences together for recognition. Liu and Shao [20] simultaneously fused the RGB and depth information using a deep architecture; Zhu et al. [58] employed a set of random forests to fuse spatiotemporal and human key joints (skeleton); Shahroudy et al. [30] selected to fuse the RGB information and skeleton cues using a structured sparsity method; [5] simply concatenated the skeleton features and silhouette-based features together for classification. However, these existing works treated the depth channel and RGB channel independently without considering their underlying connections (structures). Thus their recognition performance would often be hindered by the ignored structure learning. In this context, our model aims to jointly learn the hidden shared and specific structures among different heterogeneous features extracted from depth and color sensors, respectively. This leads to a better overall performance in the RGB-D activity recognition.

Shared-specific structures learning for activity recognition. Learning shared-specific structures for activity recognition is found to be beneficial. Shared-specific structures are defined and mined from different perspectives and for different purposes in the literature [8], [32], [37], [40], [45], [57]. For example, some researchers intended to exploit their discriminative shared-specific features by constructing shared and class-specific dictionaries [8], [37] or learning local motion patterns that are shared by different activities [57]; and recently, this idea was also introduced for recognizing activities captured from different views in [32], [45]. However, these methods assume that they can directly align different feature channels or extract shared and specific information without any pre-learning. Our proposed model differs from them significantly, since an *i-transform* is introduced for each feature channel in order to make the shared-specific structures learning be performed in a more suitable latent space. And this is highly demanded when processing heterogeneous features with different number of dimensions. Although the CCA in [3] is mostly close to ours, it is not for discriminative learning, and moreover it assumes that the specific component for each feature channel is a Gaussian distribution (or Gaussian noise) and this assumption may not hold and thus not be sufficient to describe the specific information of each channel. Our experimental results show that our JOULE model performs better than an advanced variant of CCA (MPCCA) in [3].

Heterogeneous feature learning for visual recognition. Our work is also relevant to heterogeneous features learning methods [9], [46], [53], which were mainly developed for fusing features in generic visual recognition tasks with different assumptions. Intuitively, one can develop a fusion model by concatenating features together in a standard multi-task framework without considering intrinsic connections (shared or specific structures) among features [4], [5], [58]. For instance, Cao et al. [4] built a heterogeneous feature machine (HFM) to integrate heterogeneous features with different types and different metrics for visual recognition. However, their performance is often limited by the ignored hidden connections modeling. Alternatively, some methods assume that different heterogeneous features share in the primitive feature space, a common subspace or even a common subset of input primitive features (without explicitly considering specific structures of each feature type) [9], [14], [46], [53]. For example, the work of [46] assumes that different tasks share a *common set* of input variables (i.e., a common set of input features). However, this is not the case for our RGB-D based activity recognition, since our features are of different types with different dimensionality. Among all these heterogeneous feature learning methods, the multi-task discriminant analysis (MTDA) [53] is the closest to ours. However, our model is notably different from it, even though both models unitize the concept of subspaces. MTDA assumes that there is a shared common space after projecting each type of features separately without explicitly considering the feature-specific structures. In contrast, we relax this assumption and assume that heterogeneous features are only partially related, which makes our method more applicable for describing the complex connections (shared and specific structures) among heterogeneous features extracted from RGB, depth and skeleton channels with large variations and thus obtain better recognition accuracy. In this context, we cast our model as a Frobenious-regularized least-square problem, with both *prediction* and *reconstruction* loss considered in a unified framework. This consequently leads to a better overall performance of our model in the experiments. It is worth noting that Wang et al. very recently extended the idea of jointly learning and sharing heterogeneous features and obtained the state of the art results for RGB-D object recognition [35].

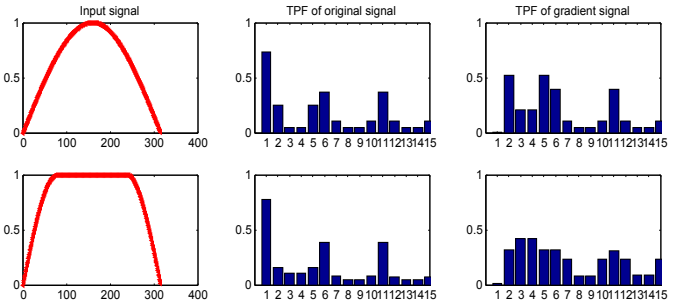


Fig. 3. Two signals (left) and their TPF features (middle and right). The TPF features of the gradient signal (right) is more distinctive than the TPF of the original signal (middle) when differentiating the input signals.

A preliminary version of this work was reported in [11]. In this paper, we have significantly extended our jointly learning framework in five aspects: 1) a new parameter was introduced to explicitly control the tradeoff between the mined shared and specific structures in the JOULE model; and more importantly 2) a new transfer learning based joint learning model was proposed by employing an auxiliary set to facilitate the feature learning on the target set; and 3) we have provided a rigorous

and theoretical analysis about the convergence of the developed three-step optimization method in the supplementary file; and 4) we conducted a new group of experiments and added more comparisons on one additional dataset consisting of a set of complex composed activities [19]; 5) we have added extensive test and deeper analysis, including the comparison with additional methods (e.g., the heterogeneous feature machine [4]), the effect of regularization parameters, the influence of the newly introduced control parameter and the evaluation of Transfer-JOULE.

3 Heterogeneous Features Construction

We describe here in detail three descriptors utilized in our model: dynamic *skeleton* (DS) features, dynamic *color* pattern (DCP) and dynamic *depth* pattern (DDP). Each descriptor consists of two components: temporal pyramid Fourier features (TPF) from: i) the original feature signal and ii) the corresponding gradient signal, respectively. These six components form our heterogeneous feature set.

The use of TPF features is motivated from the work of Wang et al. [39]. Following their practice, we repeatedly partition the feature signal (e.g., temporal skeleton features in [39]) into 1, 2 and 4 sub-segments along the temporal dimension, and then concatenate the low frequency Fourier coefficients extracted from each segment.

In addition to computing TPF from the original feature series as in [39], we also calculate TPF from the temporal gradient signal of the original feature series. This proposed extension is motivated from the following observations: 1) the gradient could, to a certain extent, implicitly encode the velocity change of the motion in activity; 2) it could also capture the variation of pixel values, which helps to describe the interactions between human and objects. For instance, the rapid change of the pixel values near a mouth may indicate that some objects are coming near and interacting with the mouth (e.g., drinking). As illustrated in Figure 3, the temporal pyramid Fourier features of the gradient signal may capture more discriminative cues.

Dynamic Skeleton. Human pose and its dynamics are one of the key elements in activities [10], [49]. Here we extract the pose dynamics using skeleton information from the depth sequences for our activity modeling. Specifically, for each video sequence, the real-time skeleton tracker [31] is used to extract the trajectories of human key joints (skeleton). Following the implementations in [39], we then compute the relative positions between each pair of trajectories and concatenate them together. The temporal pyramid Fourier features are further extracted from the relative positions as well as its gradient version to represent the dynamic pose information. It was noted that the sequence length may vary from video to video. Relative positions of each trajectory pair are interpolated by cubic spine to have the same length before computing the Fourier features, which ensures that the frequency locations of computed TPF features are properly calibrated and aligned before comparison.

Dynamic Color and Depth Pattern. Using the 3D joint positions without local appearance is often insufficient to characterize complex activities including human-object interactions. To compensate this, the local appearance features (both in RGB and depth) are extracted around each human joint, which could capture characteristic shape, texture and manipulated object's appearance during interactions. Specifically, for each joint in a trajectory, we first compute the HOG feature in its local region for all the associated frames. All of the HOG features of one joint trajectory constitute a temporal HOG tube. Then for the trajectory of each bin of the vectorized HOG feature along the time dimension, we extract the TPF features including the original and gradient version, and then concatenate them together to form our final descriptor. The HOG-TPF extracted from RGB sequence and depth sequence form our dynamic *color* pattern (DCP) and dynamic *depth* pattern (DDP), respectively.

4 Heterogeneous Feature Learning

Different features may share some similar structural components as illustrated in Figure 1. To effectively quantify the shared structures among different features with varied dimensions, we introduce a set of subspaces to represent these features so that they can be compared directly. These subspaces are learned by balancing the trade-off between the shared structures and feature-specific cues. In the following, we define our notations first, and then present a detailed description of the proposed joint learning model.

4.1 The Joint Learning Model

Suppose there are S types of heterogeneous features to learn together. For each feature type i $(i = 1, ..., S)$, let $\mathbf{X}_i \in \mathbb{R}^{d_i \times n_i}$ denote the feature matrix of n_i training instances, where d_i represents the feature dimensionality. We attempt to learn a projection matrix $\mathbf{\Theta}_i$ for each $\mathbf{X}_i$ to project it into a subspace spanned by the columns of $\mathbf{\Theta}_i$. Here for simplicity and clarity, we call this projection matrix $\mathbf{\Theta}_i$ as *intermediate transform* (i-transform).

In total, we have S subspaces, which are set to have the same dimensionality such that both the shared and feature-specific structures across different feature types can be easily quantified in the projected feature space by two weight matrices $\mathbf{W}_0, \mathbf{W}_i \in \mathbb{R}^{M \times L}$, where M is the dimensionality of the subspace, and usually $M << d_i$. L indicates the number of activity classes. We use $\mathbf{Y}_i \in \{-1, L-1\}^{L \times n_i}$ to represent the labels of all the training samples for the i^{th} feature. Each column of $\mathbf{Y}_i$ is defined as a zero-mean vector $[-1, ..., -1, L-1, -1..., -1]^T$. Note that all of the $\mathbf{Y}_i$s are label vectors and they are the same for different types of features. For a sample with class label l $(l = 1, ...L)$, the l^{th} entry of the zero-mean vector equals to a constant positive number $L - 1$.

Now, we formulate our **j**oint heter**o**geneous feat**u**res **le**arning (JOULE) model in a multi-task learning framework with orthogonality constraints considered as follows:

$$\begin{aligned}\min_{\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\}} \sum_{i=1}^{S} (\overbrace{\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2}^{R_1(\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\})} \\ + \overbrace{\gamma\|\mathbf{X}_i - \mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2}^{R_2(\{\mathbf{\Theta}_i\})} + \overbrace{\frac{\alpha}{S}\|\mathbf{W}_0\|_F^2 + \beta\|\mathbf{W}_i\|_F^2)}^{R_3(\mathbf{W}_0,\{\mathbf{W}_i\})} \\ s.t.\ \mathbf{\Theta}_i^T\mathbf{\Theta}_i = \mathbf{I}, i = 1, 2, ...S\end{aligned} \tag{1}$$

Our heterogeneous feature learning model intends to jointly learn the subspaces (encoded by i-transform $\{\mathbf{\Theta}_i\}$), shared and feature-specific components (represented by $\mathbf{W}_0$ and $\{\mathbf{W}_i\}$, respectively) in a unified framework. We cast it as a least-square problem with both prediction (the first term $R_1(\mathbf{W}_0, \{\mathbf{W}_i\}, \{\mathbf{\Theta}_i\})$) and reconstruction loss (the second term $R_2(\{\mathbf{\Theta}_i\})$) as well as the regularization term $R_3(\mathbf{W}_0, \{\mathbf{W}_i\})$

considered together. In the following, we discuss these terms in detail one by one.

Prediction loss term $R_1(\mathbf{W}_0, \{\mathbf{W}_i\}, \{\mathbf{\Theta}_i\})$. This item is defined as $(\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2)$ such that the empirical risk of each feature can be minimized, and thus it would guide our shared-specific structures learning for the purpose of better recognition. We formulate the prediction loss term in the multi-task learning framework in order to jointly learn the shared and specific structures across different features and classes together. Here, we model the structures in the weight space such that the shared-specific structures can be mined in a discriminative framework. Specifically, we use a weight matrix $\mathbf{W}_0$ that is owned jointly by different features to encode the shared structures. We also employ a matrix $\mathbf{W}_i$ only privately possessed by the i^{th} feature to capture its specific component. Discovering the shared and specific structures in a joint learning model is essential for connecting and transferring information among different tasks. Therefore our method could generalize well to the case of knowledge transfer from some auxiliary data to facilitate the model learning, which will be further elaborated in Section 5. Here, we utilize parameter $\lambda \in [0, 1]$ to control the tradeoff between the mined shared and specific structures. Larger λ leads to a larger weight on the shared structure and smaller weight on the specific structures.

Reconstruction loss term $R_2(\{\mathbf{\Theta}_i\})$. This term is defined as the reconstruction loss term to ensure that a good reconstruction (controlled by the parameter γ) can be derived from the learned subspace using i-transform during optimization, which leads to a meaningful solution of the model.

To facilitate the formulation of reconstruction loss term, an orthogonal constraint $\mathbf{\Theta}_i^T\mathbf{\Theta}_i = \mathbf{I}$ was imposed on the i-transforms. The purposes are 1) to reduce the redundancy to certain extent while preserving data information; and more importantly, 2) to establish a feasible link between points in the original and projected feature spaces. For instance, given a point $\mathbf{y} = \mathbf{\Theta_i}^T\mathbf{x}$ in the projected feature space (via $\mathbf{\Theta_i}^T$), its corresponding point in the original feature space is given by $\mathbf{\Theta}_i\mathbf{y}$; and subsequently, 3) to simplify the reconstruction loss term (shown as follows). Therefore, we can formulate the reconstruction loss as $\|\mathbf{X}_i - \mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2$.

To simplify this term further:

$$\begin{aligned}
&\|\mathbf{X}_i - \mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2 \\
=&tr(\mathbf{X}_i^T(\mathbf{I} - \mathbf{\Theta}_i\mathbf{\Theta}_i^T)(\mathbf{I} - \mathbf{\Theta}_i\mathbf{\Theta}_i^T)\mathbf{X}_i) \\
=&tr(\mathbf{X}_i^T(\mathbf{I} - \mathbf{\Theta}_i\mathbf{\Theta}_i^T)\mathbf{X}_i) \\
=&tr(\mathbf{X}_i^T\mathbf{X}_i) - tr(\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i) \\
=&\|\mathbf{X}_i\|_F^2 - \|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2
\end{aligned}$$

Here, $tr(\cdot)$ represents a matrix trace operator. By discarding the constant term $\|\mathbf{X}_i\|_F^2$, the reconstruction term can be reformulated as

$$R_2(\{\mathbf{\Theta}_i\}) = -\gamma\|\mathbf{\Theta}_i^T X_i\|_F^2. \tag{2}$$

Regularization term $R_3(\mathbf{W}_0, \{\mathbf{W}_i\})$. The regularization term $\frac{\alpha}{S}\|\mathbf{W}_0\|_F^2 + \beta\|\mathbf{W}_i\|_F^2$, a Frobenius Norm on matrices $\mathbf{W}_0$ and $\mathbf{W}_i$ parameterized by α and β (S is a constant, the number of heterogeneous features), aims to achieve a reliable generalization of our joint learning model. The two parameters can also control the values of the mined shared and specific components. Larger α leads to a smaller shared component and larger β results in smaller specific components. Integrating this regularization term can also help deriving a closed form solution of $\mathbf{W}_0$ and $\mathbf{W}_i$ during the iterative optimization presented later.

By substituting all the terms into the objective function, our problem can be rewritten as:

$$\begin{aligned}
\min_{\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\}} &\sum_{i=1,...,S}(\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2 \\
&- \gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2) + \alpha\|\mathbf{W}_0\|_F^2 + \beta\sum_{i=1,...,S}\|\mathbf{W}_i\|_F^2 \\
s.t.\ &\mathbf{\Theta}_i^T\mathbf{\Theta}_i = \mathbf{I}, i = 1, 2, ...S
\end{aligned} \tag{3}$$

4.2 Three-step Iterative Optimization

We solve our joint learning model by a coordinate descent algorithm that optimizes over one set of the parameters at each step while keeping the others fixed. The optimization is achieved by iterating the following three steps, which in a row monotonically decreases the objective function in Formula (2) with a guaranteed convergence to a local optimal solution.

STEP 1. Fixing the coefficients $\mathbf{W}_i$ and $\mathbf{\Theta}_i$, minimize the following function J_1 over $\mathbf{W}_0$:

$$\min_{\mathbf{W}_0}\sum_{i=1}^{S}\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2 + \alpha\|\mathbf{W}_0\|_F^2 \tag{4}$$

This is an unconstrained minimization problem, whose solution can be given by $\mathbf{W}_0^* = \lambda(\lambda^2\sum_i\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \alpha\mathbf{I})^{-1}\sum_i(\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}_i^T - (1-\lambda)\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i))$.

We also note that the second derivative of the objective function J_1 can be given by

$$\frac{\partial^2 J_1}{\partial\mathbf{W}_0^2} = 2(\lambda^2\sum_{i=1}^{S}\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \alpha\mathbf{I}) \succeq 0$$

where $\succeq 0$ indicates *positive semidefinite*. Hence, the derived optimal solution $\mathbf{W}_0^*$ would decrease the value of the objective function.

STEP 2. Fixing the coefficients $\mathbf{W}_0$ and $\mathbf{\Theta}_i$, optimize $\mathbf{W}_i$:

$$\min_{\{\mathbf{W}_i\}}\sum_{i=1}^{S}\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2 + \beta\|\mathbf{W}_i\|_F^2$$

The above problem can be decomposed into S independent Frobenius-regularized unconstrained least square problems:

$$\min_{\mathbf{W}_i}\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2 + \beta\|\mathbf{W}_i\|_F^2 \tag{5}$$

By setting the first order derivatives of the above function (5) to zero, we can obtain the optimal solution: $\mathbf{W}_i^* = (1-\lambda)((1-\lambda)^2\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \beta\mathbf{I})^{-1}\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}_i^T - \lambda\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_0)$. Similar to STEP 1, we can easily derive the second derivative as

$$\frac{\partial^2 J_2}{\partial\mathbf{W}_i^2} = 2((1-\lambda)^2\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i + \beta\mathbf{I}) \succeq 0$$

Here, J_2 indicates the objective function in Formula (5). Hence, it is convex with respect to $\mathbf{W}_i$, which indicates that the updating scheme at STEP 2 would decrease the value of our objective function in Formula (3) and minimize the function.

STEP 3. Finally, we fix $\mathbf{W}_0, \mathbf{W}_i$ and optimize $\mathbf{\Theta}_i$:

$$\begin{aligned}
\min_{\mathbf{\Theta}_i}&\sum_{i=1}^{S}(\|(\lambda\mathbf{W}_0 + (1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i - \mathbf{Y}_i\|_F^2 - \gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2) \\
s.t.\ &\mathbf{\Theta}_i^T\mathbf{\Theta}_i = \mathbf{I}, i = 1, 2, ...S
\end{aligned}$$

Note that all the $\mathbf{\Theta}_i$s in the above system are independent. Hence, we turn to solving the following S independent subproblems:

$$\begin{aligned}&\min_{\mathbf{\Theta}_i}\|(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i\|_F^2-\gamma\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2\\&s.t.\ \mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I}\end{aligned}\tag{6}$$

It is not easy to solve the problem in Formula (6) directly in the Euclidean space due to the non-convex constraints. We optimize each subproblem with a gradient based method on the Stiefel manifold where the approximate solution is required to satisfy the orthogonality constraint in each iteration [42]. Specifically, given the t^{th} step estimator of $\mathbf{\Theta}_i(t)$, we first define a skew-symmetric matrix $\nabla=\mathbf{G}\mathbf{\Theta}_i(t)^T-\mathbf{\Theta}_i(t)\mathbf{G}^T$, where $\mathbf{G}$ is the gradient of the objective function in the Euclidean space and it can be indicated by $\mathbf{G}=\mathbf{X}_i((\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i(t)^T\mathbf{X}_i-\mathbf{Y}_i)^T(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T-2\gamma\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i(t)$. Then the new updated point can be determined by the Grank-Nicolson-like scheme $\mathbf{\Theta}_i(t+1)=(\mathbf{I}+\frac{\tau}{2}\nabla)^{-1}(\mathbf{I}-\frac{\tau}{2}\nabla)\mathbf{\Theta}_i(t)$, where τ is the iteration step size and an optimal step size would be determined by a line search method within each iteration. We summarize the optimization for the objective function in Formula (3) in Algorithm 1.

Here, we would like to point out that the employed updating scheme at STEP 3 still makes the objective function decrease. We provide our proof in the supplementary file based on some tricks provided in [42].

As discussed above, all the three steps in our optimization method would decrease the objective function in our JOULE model. Since

$$-\|\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2=-\|\mathbf{X}_i\|_F^2+\|\mathbf{X}_i-\mathbf{\Theta}_i\mathbf{\Theta}_i^T\mathbf{X}_i\|_F^2\geq-\|\mathbf{X}_i\|_F^2,$$

the objective function in Formula (3) is lower bounded when $\alpha,\beta,\gamma\geq 0$. Therefore, the proposed optimization algorithm can converge to a minimum in practice.

4.3 Inference

Given the model parameters $\mathbf{W}_0,\mathbf{W}_i$ and $\mathbf{\Theta}_i$, the inference is to predict the best activity label for a new sample with heterogeneous features $\mathbf{x}_i, i=1,2,...,S$. We first define two confidence vectors to encode the shared and specified components of $\mathbf{x}_i$ as

$$\begin{aligned}\mathbf{C}^i_{shared}=&\lambda\mathbf{W}_0^T\mathbf{\Theta}_i^T\mathbf{x}_i\in\mathbb{R}^L\\\mathbf{C}^i_{specified}=&(1-\lambda)\mathbf{W}_i^T\mathbf{\Theta}_i^T\mathbf{x}_i\in\mathbb{R}^L\end{aligned}\tag{7}$$

Here, λ is the model parameter used to balance the contribution of the shared and specific structures during training. Specifically, when $\lambda=0$, a model without forming any shared components is formulated, while setting $\lambda=1$ formulates a baseline without specific structures explored. The effect of λ will be discussed in the experimental section.

Inspired by the construction of augmented features in [17], here we treat all the shared and specific confidence vectors as higher-level augmented features and concatenate them together to form our final representation. To speed up our testing, a linear SVM classifier was first trained on the augmented features from the training set and then subsequently used to make the final decision for a test image.

Algorithm 1 Optimization for the objective function in Formula (3). Terms $objUpdate$ and $objUpdateIn_i$ indicate the value variation of the objective function of Formula (3) and the i^{th} subproblem (6) at STEP 3, respectively.

Require:
Input: $M,\alpha,\beta,\gamma,\lambda,\mathbf{Y}_i,\mathbf{X}_i$;
Initialization: $\mathbf{W}_0,\mathbf{W}_i\in\mathbb{R}^{M\times L}$ are random matrices, $\mathbf{\Theta}_i$ is set as the top M principal components of $\mathbf{X}_i$, $IterOut=1$;
Ensure:
1: **while** $objUpdate\geq thr$ and $IterOut<maxIter$ **do**
2: $\mathbf{W}_0\leftarrow(\lambda^2\sum_i\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\alpha\mathbf{I})^{-1}$;
3: $\sum_i\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}_i^T-(1-\lambda)\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_i)$;
4: $\mathbf{W}_i\leftarrow((1-\lambda)^2\mathbf{\Theta}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i+\beta\mathbf{I})^{-1}\mathbf{\Theta}_i^T\mathbf{X}_i(\mathbf{Y}_i^T-\lambda\mathbf{X}_i^T\mathbf{\Theta}_i\mathbf{W}_0)$, i=1,2,...,S;
5: **for** i=1;$i\leq S$;$i++$ **do**
6: $IterIn=1, objUpdateIn_i=1+thr$;
7: **while** $objUpdateIn_i\geq thr$ and $IterIn\leq 50$ **do**
8: $\mathbf{G}\leftarrow\mathbf{X}_i((\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T\mathbf{\Theta}_i^T\mathbf{X}_i-\mathbf{Y}_i)(\lambda\mathbf{W}_0+(1-\lambda)\mathbf{W}_i)^T-2\gamma\mathbf{X}_i\mathbf{X}_i^T\mathbf{\Theta}_i$
9: $\nabla\leftarrow\mathbf{G}\mathbf{\Theta}_i^T-\mathbf{\Theta}_i\mathbf{G}^T$
10: $\mathbf{\Theta}_i\leftarrow(\mathbf{I}+\frac{\tau}{2}\nabla)^{-1}(\mathbf{I}-\frac{\tau}{2}\nabla)\mathbf{\Theta}_i$;
11: $IterIn++$;
12: **end while**
13: **end for**
14: $IterOut++$;
15: **end while**
16: **return** $\mathbf{W}_0,\mathbf{W}_i,\mathbf{\Theta}_i$

5 Transfer Joint Heterogeneous Feature Learning

It is challenging to learn a set of reliable i-transforms and shared-specific structures from a target set (a set where testing is carried out) with limited training samples. This situation could be mitigated by using some non-target sets (widely known as *transfer learning*). Thanks to the nature of joint learning, our JOULE model could generalize well to this case. Here, we introduce a transfer learning model to enhance our feature learning on the target set by the assistance of learning on other non-target datasets [29], [56].

Specially, we utilize samples from a non-target set as our auxiliary set to assist our feature learning on the target set and train our model on both sets in one framework. For clarity, in the following, we will use *auxiliary set* to denote non-target set. Let $\mathbf{W}_0,\{\mathbf{W}_i\}$ (and $\overline{\mathbf{W}}_0,\{\overline{\mathbf{W}}_i\}$) be the shared and specific structures to be mined in the target (and auxiliary) set, respectively. For transferring the learning from an auxiliary set to a target set, we assume that the i-transforms $\{\mathbf{\Theta}_i\}$ can be shared for the same type of features across datasets, so that the data in the auxiliary set can provide a strong prior for our feature learning on the target set. Therefore, the feature learning on the target and auxiliary sets are connected by $\{\mathbf{\Theta}_i\}$ and they can be optimized jointly. Let $\{\mathbf{X}_i^a\}$, $\{\mathbf{Y}_i^a\}$ (and $\{\mathbf{X}_i^t\}$,$\{\mathbf{Y}_i^t\}$) denote the feature representation and label information of the auxiliary set (and the training samples from target set). Our transfer joint learning model is formulated as:

$$\begin{aligned}&\min_{\substack{\mathbf{W}_0,\{\mathbf{W}_i\},\\\{\mathbf{\Theta}_i\},\overline{\mathbf{W}}_0,\{\overline{\mathbf{W}}_i\}}}\ \rho\ \underbrace{F(\{\mathbf{X}_i^t\},\{\mathbf{Y}_i^t\},\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\})}_{Learned\ on\ the\ training\ samples\ of\ target\ set}\\&+(1-\rho)\overbrace{F(\{\mathbf{X}_i^a\},\{\mathbf{Y}_i^a\},\overline{\mathbf{W}}_0,\{\overline{\mathbf{W}}_i\},\{\mathbf{\Theta}_i\})}^{Learned\ on\ the\ auxiliary\ set}\\&s.t.\ \mathbf{\Theta}_i^T\mathbf{\Theta}_i=\mathbf{I},i=1,2,...S\end{aligned}\tag{8}$$

Here, function $F(\cdot)$ is the objective function of our JOULE model (1) in the form of $\sum_{i=1}^{S}(R_1(\cdot)+R_2(\cdot)+R_3(\cdot))$. The first $F(\cdot)$ function is defined on the training samples from the target set and the second $F(\cdot)$ function on the auxiliary sets. We use the parameter $\rho \in [0,1]$ to control the effect of the auxiliary set. Specifically in the case of $\rho = 0$, the i-transforms $\{\mathbf{\Theta}_i\}$ are solely determined by the feature learning in the auxiliary set.

Similar to JOULE, we develop a three-step optimization algorithm to solve problem (8), i.e., iteratively optimizing the objective function over one set of parameters with the others fixed (e.g., at one step, we optimize over the shared components $\mathbf{W}_0$ and $\mathbf{W}_0^a$ by fixing the others.). The only difference is that the i-transforms $\{\mathbf{\Theta}_i\}$ are optimized simultaneously on both target and auxiliary datasets. The gradient of the objective function in problem (8) with respect to $\mathbf{\Theta}_i$ can be given by $\rho\frac{\partial F(\{\mathbf{X}_i^t\},\{\mathbf{Y}_i^t\},\mathbf{W}_0,\{\mathbf{W}_i\},\{\mathbf{\Theta}_i\})}{\partial \mathbf{\Theta}_i} + (1-\rho)\frac{\partial F(\{\mathbf{X}_i^a\},\{\mathbf{Y}_i^a\},\overline{\mathbf{W}}_0,\{\overline{\mathbf{W}}_i\},\{\mathbf{\Theta}_i\})}{\partial \mathbf{\Theta}_i}$, which is a combination of gradients in the target set and auxiliary set. It is easy to see that in the extreme cases when $\rho = 0$ (or $\rho = 1$), the i-transforms $\{\mathbf{\Theta}_i\}$ will be derived solely from the auxiliary set (or the target set). After all the parameters are learned, the inference step is actually identical to the JOULE model described in Section 4.3 and the corresponding decisions are made on the testing samples from the target set using the learned parameters: $\mathbf{W}_0, \{\mathbf{W}_i\}, \{\mathbf{\Theta}_i\}$.

6 EXPERIMENTS

We evaluated our methods extensively on three benchmark 3D activity datasets and one newly collected human-object interaction dataset. In the following, we first briefly introduce the implementation details, and then describe the experiments and results.

6.1 Implementation Details

The model parameters $\alpha, \beta, \gamma, \lambda$ were fixed as $10^{-1}, 10^{-1}$, 1 and $\frac{1}{2}$, respectively through all our experiments. The dimensionality M of the subspace is specified empirically for each dataset. Intuitively, it is suggested to be smaller than the number of training samples. We will investigate its effect in detail in Section 6.6. When computing DCP and DDP features, one image patch of size 60×60 was extracted around each joint position in a trajectory in order to capture the context cues. A set of image patches were extracted for each trajectory. For computational efficiency, all the image patches were then resized to 32×32 and the cell size of HOG was set to 8.

6.2 MSR Daily Activity Dataset

We tested the proposed methods on the MSR Daily Activity dataset [39], which has become a standard set for studying 3D human activities. It contains 320 video clips of 16 different activities (drinking, eating, walking, cheering up, reading book, etc) performed by 10 participants in two different poses, namely *sitting* and *standing*. Most of the activities involve human-object interactions (see Table 4). We followed the same experimental settings as in other related works, where half of the participants were used for training and the rest for testing.

To evaluate our proposed JOULE model, we compare with a baseline implementation that fuses different features together with a standard SVM classifier, MTDA [53] and HFM [4]. We denote these baselines as "SVM", MTDA, and HFM. In addition, we also compare with the MPCCA model presented in [3], which intends to discover shared-specific structures in a non-discriminative learning framework. We also present the recently reported results of other 10 different methods for comparison. The dimensionality M for our JOULE model was set to 40.

Results. Table 1 shows the results and comparison. Our method obtains an accuracy of 95%, which exceeds most of the latest reported results and is comparable with the state-of-the-art [21]. However, we would like to point out that Lu et al. [21] requires a clear pixel-wise segmentation of the actor, background and occlusion objects, which may render it unsuitable for activities with more complex interactions and cluttered background. Compared to the closely related methods focusing on feature fusion using deep model [20] and structured sparse model [30], our model outperforms both of them by a considerable margin (more than 9.4%), which implies our feature learning system is superior to other RGB-D activity fusion systems. Compared with the baseline of SVM, the performance gain (95% vs. 90%) by our JOULE model demonstrates the benefits of the shared and specific components modeling. Our JOULE outperforms MTDA and HFM considerably by 4.4% and 10.6% using exactly the same set of features. It is worth noting that MTDA did not seek to learning feature-specific structures. The superior performance of our JOULE over MTDA indicates that modeling feature-specific structures is essential for capturing the complex connections among the employed heterogeneous features. It is also observed that HFM performed worse than the baseline of SVM. Bear in mind that, in order to compute the similarity of two training instances, HFM needs to manually select a proper kernel for each feature type, which is a big challenge in the presence of noisy heterogeneous features (e.g., part of our DCP features were extracted from the background pixels). Therefore, the resulting similarity matrix could be unreliable and HFM might not cope with our features well. In our implementation of HFM, we used both RBF kernel and linear kernel to measure the similarity between two features, which was suggested in [4]. In contrast, the SVM will adaptively learn a set of weights to encode the contribution of each feature dimension in a discriminative framework and thus can be more applicable in our RGB-D activity recognition. The MPCCA is an approach close to the proposed JOULE, but it performed clearly inferior to JOULE. One of the reasons is that the Gaussian noise assumption in MPCCA is not sufficient to describe the specific information of each feature channel. Moreover, JOULE also benefited from learning discriminant shared-specific structure.

The confusion matrix of the results by our JOULE model is shown in Figure 4. It can be seen that our model achieves perfect classification results on 10 classes. The larger error is due to the mis-classification of the activity of *writing on a paper* as *reading book*, which may be largely attributed to high similarity between the object and activity contexts in these two activities.

6.3 Cornell Activity Dataset 60 (CAD 60)

This public dataset consists of 68 video clips captured by Microsoft Kinect device [33]. Four actors were asked to perform 13 specific activities (*still*, *talking on the phone*, and etc.) and one random activity in 5 different environments: office, kitchen, bedroom, bath room, and living room. We followed the same experimental setting in [39] by adopting the leave-one-person-out cross validation for each environment, which ensures that person participating in the training cannot be seen in the testing. The final

TABLE 1
Comparison on the MSR Daily Activity dataset.

	Method	Accuracy(%)
Reported Results	Dynamic Temporal Warping [25]	54
	3D Joints and LOP Fourier [39]	78
	HON4D [28]	80.00
	SSFF [30]	81.9
	Deep Model (RGGP) [20]	85.6
	Actionlet Ensemble [39]	85.75
	Super Normal [48]	86.25
	Bilinear [14]	86.88
	DCSF+Joint [43]	88.2
	LFF+IFV [51]	91.1
	Group Sparsity [22]	95
	Range Sample [21]	**95.6**
Our Results	HFM [4]	84.38
	SVM	90
	MPCCA [3]	90.62
	MTDA [53]	90.62
	JOULE	**95**

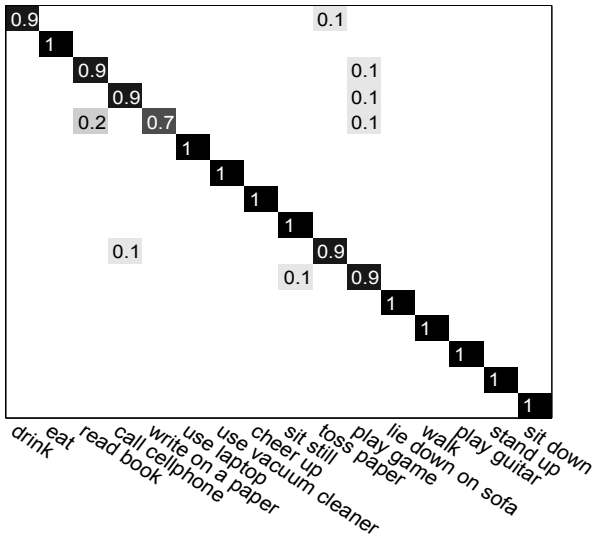


Fig. 4. Confusion matrix of JOULE on MSR Daily dataset.

accuracy was calculated by averaging the accuracies of all the possible splits (totally 20 in this set).

Our methods are compared with the results reported in the state-of-the-art [39]. We also ran the released code of HON4D on this set and listed the recognition results as "Reported Results" in Table 2. Since there is no default parameter settings suggested by the author on this set, we report the best results by varying their parameters in a wide range. Similar to MSR Daily set, we also highlight the benefits of using JOULE model by comparing with the baseline SVM, MTDA and HFM. Here, the dimensionality M of $\mathbf{W}_i$ (and $\mathbf{W}_0$) is set as 4 on this dataset.

Results. The results and comparison are shown in Table 2. Our method achieves an accuracy of 84.1%, which significantly outperforms the state-of-the-art result [39] by a large margin (9.4%). It is worth noting that most of our baseline implementations including the simple combination of our heterogeneous features with a standard SVM classifier can achieve a performance comparable to the state-of-the-art method with carefully designed classifiers, which proves that our feature is superior to that developed in [39]. Especially, by considering the shared and specific components,

TABLE 2
Comparison on the CAD 60 dataset.

	Method	Accuracy(%)
Reported Results	STIP [59]	62.5
	Order Sparse Coding [26]	65.3
	Object Affordance [15]	71.4
	HON4D [28]	72.7
	Actionlet Ensemble [39]	74.7
Our Results	HFM [4]	72.7
	SVM	75
	MPCCA [3]	79.1
	MTDA [53]	82.6
	JOULE	**84.1**

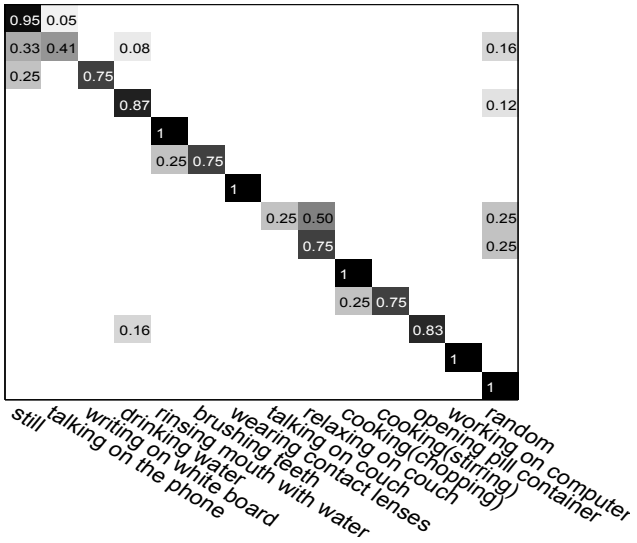


Fig. 5. Confusion matrix of JOULE on CAD 60 set.

our model (JOULE) obtains a gain of 9.1% compared with the fusion methods using standard SVM classifier without explicitly modeling shared and specific components (84.1% vs. 75%), and a significant gain of 11.4% compared with HFM. In addition, our JOULE works better than MTDA on CAD 60 set with a smaller performance gain than on the MSR Daily set.

The confusion matrix of the results by our JOULE model is presented in Figure 5. It can be seen that our model can distinguish well the five activities of rinsing mouth with water, wearing contact lenses, cooking(chopping), working on computer and random activities, which demonstrates that our model can effectively capture the interactions between human and the manipulated object. It can also be observed that the activities of talking on couch and relaxing on couch are often confused by our model, mainly due to the inaccurate human skeletons captured by the Kinect camera.

6.4 Composable Activities Dataset

This dataset consists of 693 video clips performed by 14 participants [1]. Each participant was asked to perform 16 complex activities (*Walk while calling with hands*, *Walk while hand waving*, and etc.) several times. All the considered activities in this set are composed by a number of mid-level actions such as walking, waving hand, reading etc., and about 75% of them contain human-object interactions. For a fair comparison, we followed exactly the

1. http://web.ing.puc.cl/~ialillo/ActionsCVPR2014/

TABLE 3
Comparison on the Composable Activities dataset.

	Method	Accuracy(%)
Reported Results	HON4D [28]	83.29
	Hierarchical Model [19]	85.7
Our Results	HFM [4]	84.44
	SVM	88.32
	MPCCA [3]	90.76
	MTDA [53]	92.07
	JOULE	**94.24**

same leave-one-subject-out experimental setting as in [19], where each time the activity samples performed by 13 participants were all used to train a model and the rest were used for testing. And finally, the average accuracies were computed and reported.

Here, we directly compare the performance of our method with the results reported in the state-of-the-art [19]. Meanwhile, we also ran the released code of HON4D by the author on this set, and again report the best results by varying their parameters in a wide range. In addition, we further compared the JOULE with the baseline "SVM", MTDA and HFM. In this experiment, we set the dimensionality M of the subspace as 100. Its influence would be further discussed in Section 6.6.

Results. The results and comparison are shown in Table 3. As shown, simply feeding the concatenation of all primal heterogeneous features into a SVM classifier without explicitly considering their hidden structures and connections achieves an accuracy of 88.32% and outperforms the state-of-the-art [19] by a margin of 2.6%. As expected, the performance gap becomes larger ($\geq 5.9\%$) when our proposed JOULE model is employed to explicitly model the shared and specific structures among different heterogeneous features. Similar to the observations on other datasets, our JOULE outperforms MTDA by over 2% on the Composed Activities Datasets, which once again experimentally confirms that the learning of feature-specific structures is beneficial.

By closely examining the confusion matrix in Figure 6, we can observe that JOULE achieves perfect recognition performance on most of the activities. The most challenging activities for our model are "*Walk while calling with hands*" and "*Walk while hand waving*", which are often confused with each other. This is not surprising, because these two activities contain highly similar motions, and the subtle difference between them is that activity "*calling with hands*" often involves a motion of moving fingers or hands back and forth, while "*waving hands*" refers to a slight hand movement of moving between left and right. However, it is quite challenging to capture these tiny differences by the prevailing Kinect cameras available in the market with standard specification of spatial and depth resolution.

6.5 SYSU 3D Human-Object Interaction Set

Dataset Description. We have collected a new RGB-D activity dataset focusing on human-object interactions to further evaluate all methods. We name this as *SYSU 3D Human-Object Interaction* (HOI) dataset. For building this set, 40 participants were asked to perform 12 different activities freely. For each activity, each participant manipulates one of the six different objects: phone, chair, bag, wallet, mop and besom. Therefore, there are totally 480 video clips collected in this set. The contained activity samples have different durations, ranging from 1.9s to 21s. For each video

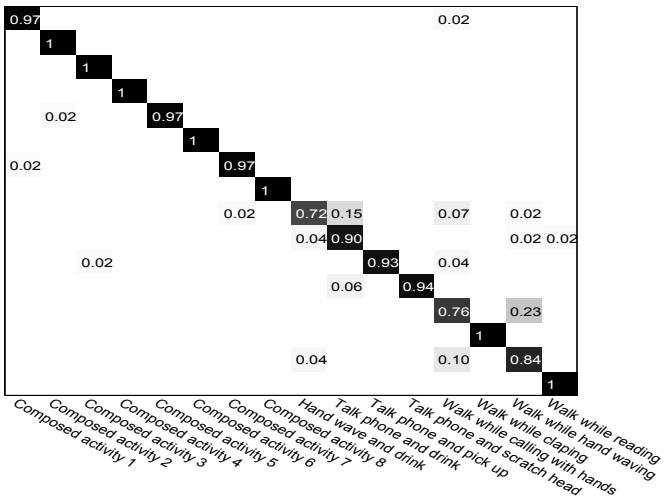


Fig. 6. Confusion matrix of JOULE on Composable Activities dataset.

TABLE 4
Comparison of 3D HOI dataset with relevant datasets. Cla. denotes *class*, and Sub. for *subject*, Vid for *video*, HOI Ra. for HOI *ratio* among the dataset.

DataSet	Data	Cla. No.	Sub. No.	Vid. No.	HOI Ra.
CAD 60 [33]	RGB-D	14	4	68	85.7%
MSRDaily [39]	RGB-D	16	10	320	87.5%
MSRAction [18]	Depth	20	10	567	$\leq 70\%$
Comp. Activities [19]	RGB-D	16	14	693	75%
Multiview [41]	RGB-D	8	8	3815	100%
SYSU 3D HOI	RGB-D	12	40	480	100%

clip, the corresponding RGB frames, depth sequence and skeleton data were captured by a Kinect camera. Activity samples are shown in Figure 8. We highlight the differences between our 3D HOI set and relevant existing sets in Table 4. Compared to those datasets (MSRDaily, CAD 60, MSRAction, Composable Activities dataset, and Multiview set), our dataset presents new challenges: 1) the involved motions and the manipulated objects' appearance are highly similar among some activities; for instance, the manipulated objects besom and mop involved in the activities mopping and sweeping are highly similar; 2) the number of participants is three times (or even larger than in most cases) that of existing ones, so that more inter-subject variations could be observed for the same type of activities due to the different characteristics of participants.

Evaluation Protocol. We tested all the compared methods in two different settings. In the first setting (setting-1), for each activity class, we selected half of the samples for training and the rest for testing. In the second setting (setting-2), video sequences performed by half of the participants were used to learn model parameters and the rest for testing, where there is no overlap of participants between the training and test set. This is a *cross-subject* setting. For each setting, we report the mean accuracy and standard deviation of the results over 30 random splits.

Baselines. Similar to that on the MSRDaily and CAD60 sets, the baselines SVM, HFM, MPCCA, MTDA and HON4D are compared to show the effectiveness of our joint learning model (JOULE). We set $M = 30$ in our model. In total, we report a comprehensive set of results of up to *six* different implementations

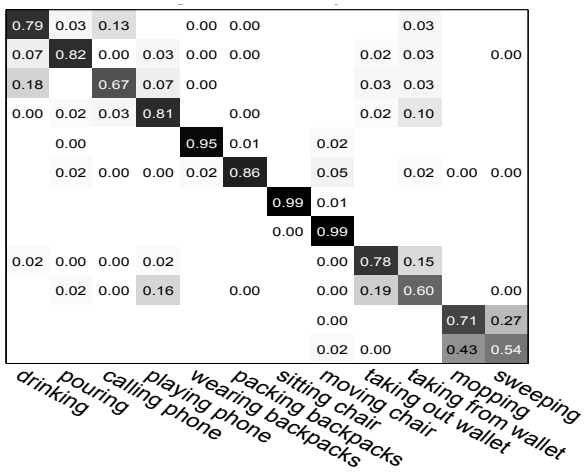
TABLE 5
Comparison on the SYSU 3D HOI dataset.

Method	Mean Acc±std (%)	
	setting-1	setting-2
HON4D [28]	73.39 (±2.59)	79.22 (±2.36)
HFM [4]	75.03 (±2.68)	76.74 (±2.63)
MPCCA [3]	76.25 (±2.36)	80.72 (±2.07)
SVM	77.34 (±2.53)	82.78 (±2.83)
MTDA [53]	79.19 (±4.27)	84.21 (±2.19)
JOULE	**79.63 (±2.13)**	**84.89 (±2.29)**

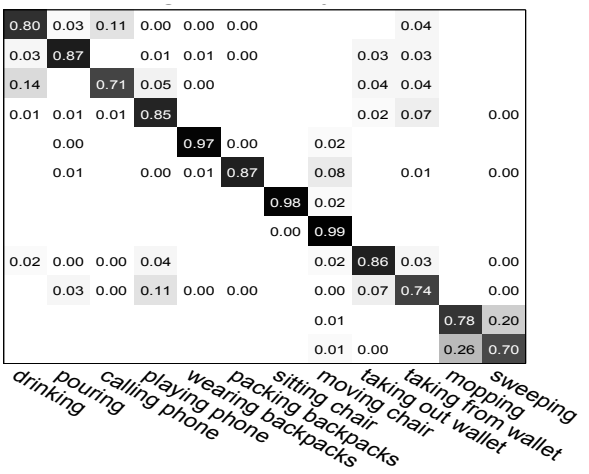
on this new dataset.

Results. Table 5 reported the results. Again, using the proposed JOULE model to fuse different heterogeneous features is always beneficial in all settings. The accuracies in setting-2 are higher than that of setting-1 without considering cross-subject split. This is because the prediction could be biased by appearance when activities with similar motion and object context (e.g. mopping vs. sweeping) performed by the same participant are contained in both training and test sets, which may occur in the setting-1. The performances of JOULE and MTDA are comparable with JOULE performing perceivably better. It was noted that the performance gap between our models and the baselines is smaller (e.g., 84.9% vs. 82.8%) than that on the other three datasets. This somehow indicates the new dataset is more challenging for feature fusion.

By examining the confusion matrices of our JOULE model in Figure 7, we observed that our model often confuses the activities of mopping with sweeping in both settings, which is mainly due to similar motions and objects appearance in the two interactions. In addition, the activities of *taking from wallet* share similar motions with activities of *playing phone* and *taking out wallet*, which are occasionally misidentified as *playing phone* or *taking out wallet*.



(a) setting-1



(b) setting-2

Fig. 7. Confusion matrices of JOULE on SYSU 3D HOI set under setting-1 (a) and setting-2 (b).

6.6 Analysis and Discussion

Convergence. Our method converges to a minimum after a limited number of iterations. We empirically observed that 20 iterations (outer iterations i.e. term $IterOut$ in Algorithm 1) are sufficient for obtaining a reliable solution in all of our experiments. See Figure 9 for an example illustrating the convergence of our method on the MSR Daily activity set, where the objective value of each step was recorded during each iteration. Excluding the time for computing the features, one round training of our algorithm takes about 1.26 minutes per training sample. However, our testing is pretty fast, and takes about 0.5 second per sample. Computing the DS, DCP, and DDP features costs time. It takes about 0.24 second for processing each frame of a RGB-D video using MATLAB on a normal desktop PC (CPU i5-4570, memory 28G).

Effect of dimensionality M. We investigate the effect of the dimensionality M of the subspace. Figure 10 shows the performances of our method JOULE with different values of M. Generally, a very small M leads to an inferior performance, as the smaller dimensionality of the subspace is, the less representative it is for the original features. When M becomes larger (typically larger than a value about $\frac{1}{6} \sim \frac{1}{4}$ of the number of training samples), the performances start to remain stable, which means our algorithm is not sensitive M in a reasonable range.

Effect of TPF on gradient signal. In this work, we have modified temporal pyramid Fourier features (TPF) developed in [39] so as to apply both the original feature signal and its gradient to implicitly

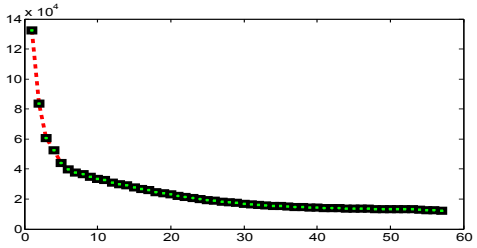


Fig. 9. Illustration of the convergence of our method. The vertical axis indicates the value of objective function and the horizontal axis is the number of iterations.

encode human motions, since they are complementary to each other. The TPF of original signal captures the original signal cues, whereas the TPF of gradient signal encodes the first derivative (velocity) information. Table 6 shows the results of our model with and without temporal Fourier features computed from the gradient signal on all of the three datasets. It can be seen that, while the improvement on the SYSU 3D HOI dataset is relatively mild, TPF features on gradient consistently improve the results in all of the cases, with the biggest gain (7.6%) achieved on the CAD60 dataset. This indicates that the proposed extension of TPF features to the gradient signal is promising and effective.

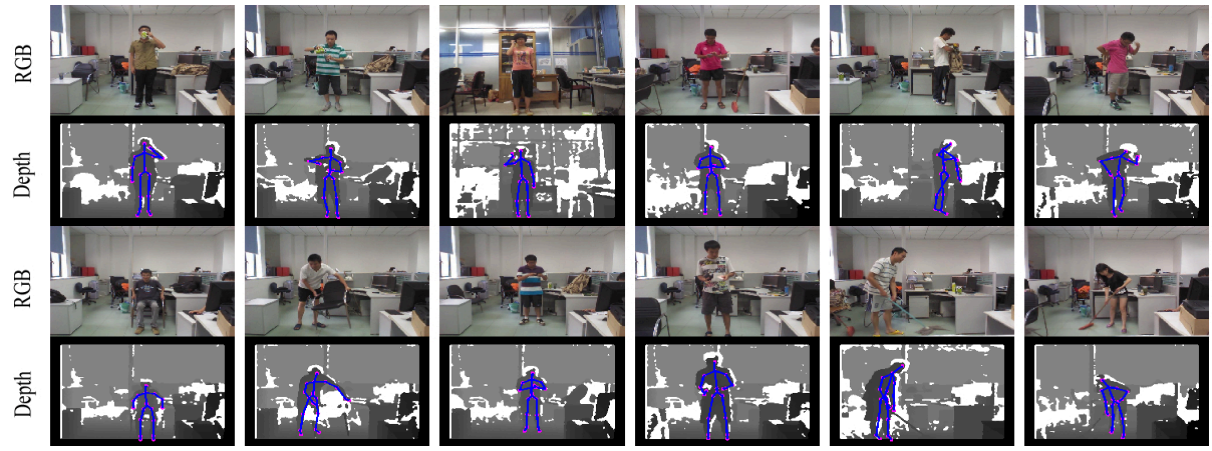


Fig. 8. Snapshots of activities in SYSU 3D HOI set, one sample per class. The rows headed with *RGB* show the samples in RGB channel and the rows underneath headed with *Depth* show the corresponding depth channel superimposed with skeleton data. Best viewed in color.

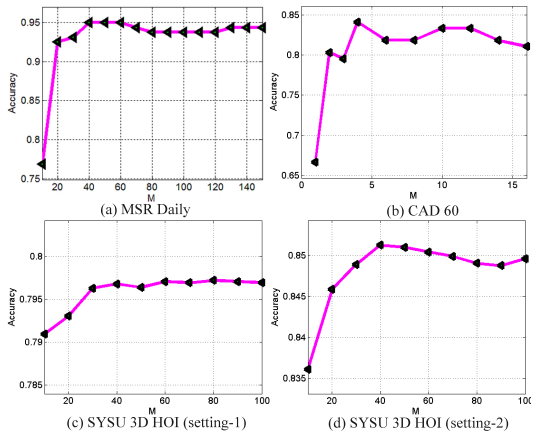


Fig. 10. Effects of parameter M on the system performance.

TABLE 6
Accuracy (%) of our methods with and without TPF on gradient. s-1 denotes setting-1 and s-2 for setting-2 applied on the SYSU 3D HOI dataset.

	MSRD	CAD60	Comp. Act.	3DHOI(s-1)	3DHOI(s-2)
With	95	84.1	94.24	79.63	84.89
Without gradient	91.25	76.5	92.22	78.83	83.63

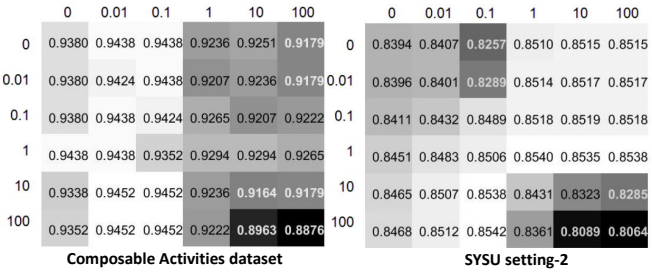


Fig. 11. Effects of parameters α (the vertical) and β (the horizontal) on the system performance (%) on the cross-subject settings of Composable Activities Dataset and SYSU set.

TABLE 7
Effects of parameter λ on recognition (%).

Dataset	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
MSRD	90.62	91.87	**95**	92.5	91.25
CAD60	82.58	83.33	84.1	**85.61**	82.58
Comp. Act.	91.21	93.80	**94.24**	93.37	91.50
SYSU(s-1)	78.83	**80.24**	79.63	79.50	78.89
SYSU(s-2)	83.81	84.65	84.89	**85.15**	84.58
Average	85.41	86.78	**87.57**	87.23	85.76

Effect of α and β. As discussed in previous sections, the parameters α and β were employed to control the generalization ability of our joint learning model. Here, we investigate their influence on Composable Activities dataset and SYSU with setting-2, where cross-subject settings (i.e., half of the subjects for training, and the rest for testing) are employed. In this test, parameters α and β were both selected from $\{0, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, and therefore we have a total of 36 different parameter settings. We present the recognition results in Fig. 11. It could be observed that, generally large α and β (≥ 10) lead to an inferior performance. This is because the larger the α and β are, the less the shared and specific components are discovered for recognition. However, when α and β are smaller than 1, the performance would remain relatively stable in most cases, which demonstrates that our method is insensitive to the parameters in a reasonable range. This study also reveals that the optimal ranges of α and β are approximately the same, which indicates that we can simply set $\alpha = \beta$ (e.g., both were set as 0.1 in all of the other experiments) to reduce the number of parameters without affecting the system performance too much.

Influence of λ. In our joint learning framework, we introduce a parameter λ to explicitly control the trade-off between the shared structure $\mathbf{W}_0$ and feature-specific structures $\{\mathbf{W}_i\}_{i=1,2,\ldots,S}$. Here, we evaluate its influence by setting λ as 0, 0.25, 0.5, 0.75, and 1, respectively, and then report the achieved performances in Table 7. As expected, a proper combination of the shared and specific structures gives a better result; generally too small or too large λ would result in an inferior performance. Especially, without modeling the specific structures ($\lambda = 1$) or shared structure ($\lambda = 0$), the performance decreased in both cases. Overall, albeit not always the best, on all of the four datasets considered, $\lambda = 0.5$ is an acceptable setting.

TABLE 8
Effects of parameter γ on recognition accuracy (%).

Dataset	$\gamma = 0$	$\gamma = 0.01$	$\gamma = 1$	$\gamma = 100$	$\gamma = 10000$
MSRD	90.6	91.5	95	93.8	90
CAD60	79.6	81.1	84.1	78.0	76.5
Comp. Act.	91.4	93.4	94.2	93. 7	92.80
SYSU(s-1)	77.0	80.0	79.6	76.8	76.9
SYSU(s-2)	81.6	83.0	84.9	84.1	82.9

TABLE 9
Effects of jointly learning in different channels. s-1 denotes setting-1 and s-2 for setting-2 applied on the SYSU 3D HOI dataset (%).

Data Channel	MSRD	CAD60	Comp. Act.	SYSU(s-1)	SYSU(s-2)
RGB	86.9	78.0	88.9	71.6	80.0
DEP	84.4	79.6	88.3	74.3	82.3
SKL	75	77.9	91.2	75.5	76.9
DEP+RGB	87.5	80.3	90.1	74.8	82.6
RGB+SKL	91.3	81.1	93.2	76.9	81.4
DEP+SKL	90.6	82.6	93.2	79.7	83.5
DEP+RGB+SKL	**95**	**84.1**	**94.2**	**80.2**	**84.9**

Influence of γ. In the JOULE model (Formula 1), we employed a reconstruction loss term (parametered by γ) to regularize the i-transforms learning in order to preserve as much information as possible. Here, we investigate its influence by varying it systemically. The results are presented in Table 8. As shown, the model performed the best when $\gamma = 1$ on most of the datasets. In general, a smaller or larger γ would lead to lower recognition accuracies. In particular, when γ is zero and the reconstruction term is not used to constrain the i-transforms learning, lower recognition results were observed.

Single vs. Multi Channels. In the JOULE model, we have integrated the learning of features from different channels (RGB, depth (DEP) and skeleton (SKL)) in a framework so that the learning of one channel can facilitate the learning of other channels. To investigate the benefits of joint learning, we tested the JOULE by feeding it with 1) features from one channel only and 2) features from two or more channels, respectively. Therefore, we tested 7 cases for each dataset. In total we conducted 35 experiments, and results are summarized in Table 9. It can be seen that the performances of learning features from two channels are higher than each of them alone. Using features from three channels always outperform one or two channels. This demonstrates that jointly learning the features from different channels is beneficial.

6.7 Experiments on Transfer-JOULE

In this section, we tested the performance of Transfer-JOULE (Formula (8)) and show how the auxiliary set can benefit our heterogeneous features learning on the target set. The experiments were carried out on the SYSU 3DHOI and Composable Activities sets as they are the two largest datasets among those considered.

Firstly, we evaluated the effect of the control parameter ρ by varying its value from 0 to 1. In this evaluation, one of the two datasets is considered as a target set, and the other as the auxiliary set. When SYSU 3DHOI was used as the target set, we followed two different settings (setting-1 and setting-2) as in Section 6.5. When Composable Activities dataset was used as the target set, we followed the leave-one-subject-out setting as described in 6.4. Thus in total, we have three different test cases: 1) SYSU 3DHOI $\rightarrow$ Composable Activities dataset; 2) Composable Activities dataset $\rightarrow$ SYSU 3DHOI (setting-1); 3 Composable Activities dataset $\rightarrow$ SYSU 3DHOI (setting-2), where $\rightarrow$ indicates the direction of transfer, i.e. auxiliary set $\rightarrow$ target set. In each case, we employed the same evaluation protocol as that in section 6 by reporting the average accuracy over a number of different training/test splits (i.e., 14 in Composable Activities dataset and 30 in SYSU 3DHOI set) on the target set. To illustrate the effectiveness of the proposed transfer learning framework, we also implemented a baseline that directly trains a JOULE (Formula 3) model on the pooled dataset that contains both the training set (from the target set) and the entire auxiliary set. This is a naive case denoted as "ExTrain".

TABLE 10
Comparison of Transfer-JOULE and JOULE, and the effects of ρ, where $\rightarrow$ indicates the direction of transfer (%).

Dataset	JOULE	Transfer-JOULE				ExTrain
	$\rho = 1$	$\rho = 0.8$	$\rho = 0.6$	$\rho = 0.4$	$\rho = 0$	
SYSU $\rightarrow$ Comp.	94.24	**95.10**	94.81	92.80	92.07	91.93
Comp. $\rightarrow$ SYSU(s-1)	79.63	80.10	**80.71**	79.54	78.58	77.19
Comp. $\rightarrow$ SYSU(s-2)	84.89	84.92	**85.15**	84.51	81.14	81.11

The results are summarized in Table 10. As shown, a proper combination ($\rho \geq 0.6$) of the feature learning in target set and auxiliary set usually improves the recognition accuracy compared to the performance of using target training set only ($\rho = 1$). The performance decreases when ρ is getting smaller. In general, setting $\rho = 0.6$ produces the best overall performance. It is observed that the direct use of i-transforms learned on auxiliary set ($\rho = 0$) can also result in a good performance on the target set, which indicates that the i-transforms could generalize well from one to the other. The superior performance of "Transfer-JOULE" over "ExTrain" shows the better capability of Transfer-JOULE in transferring information gained in auxiliary set to target set. Note that the Transfer-JOULE always performs better than the (non-transfer) JOULE trained on the pooled dataset. This suggests that simply merging the auxiliary and target datasets together is not an optimal way to exploit the transferrable shared-specific structures.

Finally, we investigate the influence of the number of the training samples in the target set. Here, we compare the performances of our JOULE model *with* and *without* transfer learning (i.e. Transfer-JOULE (8) and JOULE (1)). As suggested in the last experiment, the parameter ρ for Transfer-JOULE is set as 0.6. The methods are evaluated when the SYSU-3DHOI set is used as the target set under two different settings (setting-A and setting-B). In setting-A, we randomly selected a certain number of samples per class to train the model and used the rest for testing. In setting-B, we randomly selected a certain number of participants and used all the samples performed by them as the training set. For a fixed number of training samples (or participants) in each setting, we report the average accuracy obtained by 30 trials. The Composable Activities dataset is used as the auxiliary set in both settings. The results are presented in Figure 12. It is observed that in all of the cases tested, with the help of auxiliary set, the performance of our Transfer-JOULE model is *consistently* higher than that of JOULE. When the number of training samples is smaller (e.g., less than 15), the performance gap gets much larger. The performance gain of using the auxiliary set becomes smaller but clearly noticeable when the number of training samples gets larger, which is as expected. In particular, in the case of one-shot

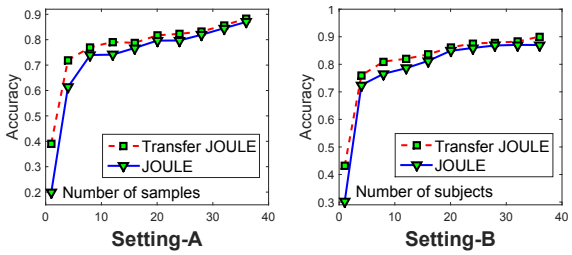


Fig. 12. Effects of the number of training samples per class (left - setting-A), and the number of subjects (right - setting-B) when the SYSU set is used as the target set.

activity recognition where only one target training sample per class is available for the model training, our Transfer-JOULE model can obtain accuracies of 39.17% and 43.04% in the setting-A and setting-B, respectively, which are about 13% higher than the (non-transferred) JOULE model. This clearly demonstrates that with the help of an auxiliary set, our Transfer-JOULE model can learn a set of parameters with better generalization than the (non-transferred) JOULE model.

7 Conclusion

We have proposed a new RGB-D method called **j**oint hete**r**ogenous feat**u**res **le**arning (JOULE) model to jointly learn heterogeneous features with different number of dimensions for RGB-D activity recognition. A transfer version is also introduced to further facilitate the joint learning on target set via exploiting shared intermediate transforms (i-transforms) from non-target data. Extensive results are reported on four RGB-D activity sets, demonstrating the effectiveness of the proposed methods. A limitation of our method is the assumption that all the considered activities should be fully executed and observed by the system, which makes it less applicable for identifying ongoing activities containing partial activity execution. In the future, we would like to extend the JOULE model so that it can be used for real-time activity recognition or prediction.

Acknowledgment

The authors would like to thank all reviewers' constructive advice for improving this work. This work was supported partially by the National Key Research and Development Program of China(2016YFB1001002, 2016YFB1001003), NSFC (61573387,61472456, 61522115, 61661130157, 61628212), Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, the GuangDong Program (2015B010105005), the Guangdong Science and Technology Planning Project (2016A010102012,2014B010118003), and Guangdong Program for Support of Top-notch Young Professionals (2014TQ01X779). The corresponding author for this paper is Wei-Shi Zheng.

References

[1] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *International Conference on Machine Learning*, pages 17–24. ACM, 2007.

[2] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, (6):1817–1853, 2005.

[3] Z. Cai, L. Wang, and X. P. Y. Qiao. Multi-view super vector for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 596 – 603, 2014.

[4] L. Cao, J. Luo, F. Liang, and T. S. Huang. Heterogeneous feature machines for visual recognition. In *IEEE International Conference on Computer Vision*, pages 1095–1102, 2009.

[5] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *IEEE International Conference on Computer Vision Workshops*, pages 91–97, 2013.

[6] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning a shared predictive structure from multiple tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1025–1038, 2013.

[7] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[8] T. Guha and R. K. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1576–1588, 2012.

[9] S. Han, X. Liao, and L. Carin. Cross-domain multitask learning with latent probit models. *arXiv preprint arXiv:1206.6419*, 2012.

[10] J. Hu, W. Zheng, J. Lai, S. Gong, and T. Xiang. Exemplar-based recognition of human-object interactions. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):647–660, 2016.

[11] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5344–5352, 2015.

[12] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, and J. Lai. Real-time rgb-d activity prediction by soft regression. In *European Conference on Computer Vision*, pages 280–296. Springer, 2016.

[13] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *International Joint Conference on Artificial Intelligence*, pages 2466–2472, 2013.

[14] Y. Kong and Y. Fu. Bilinear heterogeneous information machine for rgb-d action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1054–1062, 2015.

[15] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[16] J. Lei, X. Ren, and D. Fox. Fine-grained kitchen activity recognition using rgb-d. In *ACM Conference on Ubiquitous Computing*, pages 208–211. ACM, 2012.

[17] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148, 2014.

[18] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–14, 2010.

[19] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 812–819, 2014.

[20] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *International Joint Conference on Artificial Intelligence*, pages 1493–1500, 2013.

[21] C. Lu, J. Jia, and C.-K. Tang. Range-sample depth feature for action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 772–779, 2014.

[22] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE International Conference on Computer Vision*, pages 1809–1816, 2013.

[23] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European conference on computer vision*, pages 359–372, 2006.

[24] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision*, pages 104–111, 2009.

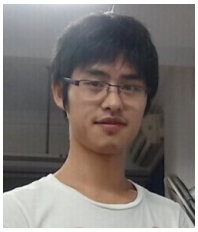
[25] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 137–146, 2006.

[26] B. Ni, P. Moulin, and S. Yan. Order-preserving sparse coding for sequence classification. In *Europeon Conference on Computer Vision*, pages 173–187, 2012.

[27] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1):24–38, 2014.

[28] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 716–

723, 2013.
[29] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
[30] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *International Symposium on Control, Communications, and Signal Processing*, pages 1–4, May 2014.
[31] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
[32] Y. Song, L.-P. Morency, and R. Davis. Multi-view latent variable discriminative models for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2120–2127, 2012.
[33] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *Plan, Activity, and Intent Recognition*, 64, 2011.
[34] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.
[35] A. Wang, J. Cai, J. Lu, and T.-J. Cham. Mmss: multi-modal sharable and specific feature learning for rgb-d object recognition. In *IEEE International Conference on Computer Vision*, pages 1125–1133, 2015.
[36] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.
[37] H. Wang, C. Yuan, W. Hu, and C. Sun. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognition*, 45(11):3902–3911, 2012.
[38] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *Europeon Conference on Computer Vision*, pages 872–885. 2012.
[39] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5), 2014.
[40] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann. Action recognition by exploring data distribution and feature correlation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1370–1377, 2012.
[41] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4d human-object interactions for event and object recognition. In *IEEE International Conference on Computer Vision*, pages 3272–3279, 2013.
[42] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.
[43] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.
[44] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, 2012.
[45] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe. Multitask linear discriminant analysis for view invariant action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.
[46] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing Systems*, pages 2151–2159, 2009.
[47] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 14–19, 2012.
[48] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 804–811, 2014.
[49] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5d graph matching. In *Europeon Conference on Computer Vision*, pages 173–186, 2012.
[50] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012.
[51] M. Yu, L. Liu, and L. Shao. Structure-preserving binary representations for rgb-d action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2015.
[52] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.
[53] Y. Zhang and D.-Y. Yeung. Multi-task learning in heterogeneous feature spaces. In *Conference on Artificial Intelligence*, 2011.
[54] Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012.
[55] Y. Zhao, Z. Liu, L. Yang, and H. Cheng. Combing rgb and depth map features for human activity recognition. In *IEEE Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–4, 2012.
[56] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2016.
[57] Q. Zhou, G. Wang, K. Jia, and Q. Zhao. Learning to share latent tasks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2264–2271, 2013.
[58] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 486–491, 2013.
[59] Y. Zhu, W. Chen, and G. Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 32(8):453–464, 2014.



Jian-Fang Hu received the PhD and B.S. degrees from the School of Mathematics, Sun Yat-Sen University, Guangzhou, China, in 2016 and 2010, respectively. His research interests include human-object interaction modeling, 3D face modeling, and RGB-D activity recognition. He has published several scientific papers in the international journals and conferences including IEEE TPAMI, IEEE TCSVT, PR, ICCV, CVPR, and ECCV.

Wei-Shi Zheng received the PhD degree in applied mathematics from Sun Yat-Sen University in 2008. He is a professor in Sun Yat-sen University. He has been a postdoctoral researcher on the EU FP7 SAMURAI Project at Queen Mary University of London. His research interests include person/object association and recognition in visual surveillance. He is a recipient of excellent young scientists fund of the national natural science foundation of China. He has joined Microsoft Research Asia Young Faculty Visiting Programme.

Jian-Huang Lai received the Ph.D. in mathematics from Sun Yat-Sen University in 1999. He is a Professor and the Dean of the School of Information Science and Technology. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet, and its applications. He has published over 100 scientific papers in international journals and conferences including IEEE TPAMI, IEEE TNN, IEEE TIP, IEEE TSMC-B, PR, ICCV, CVPR, and ICDM.

Jianguo Zhang is currently a Reader of Visual Computation at University of Dundee, UK. He received a PhD in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2002. His research interests include visual surveillance, object recognition, image processing, medical image analysis and machine learning.